Voice Agent 모델 논의

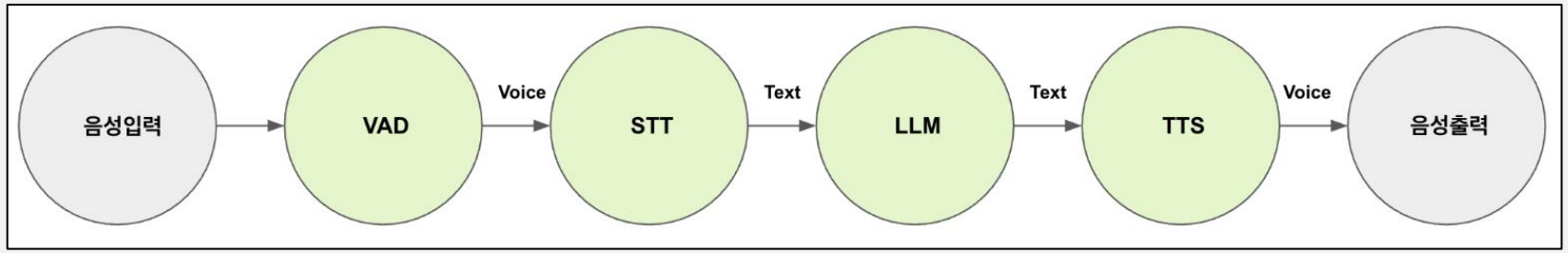
2025.06.17

에이닷사업부

I. 에이닷 Voice Agent 현황

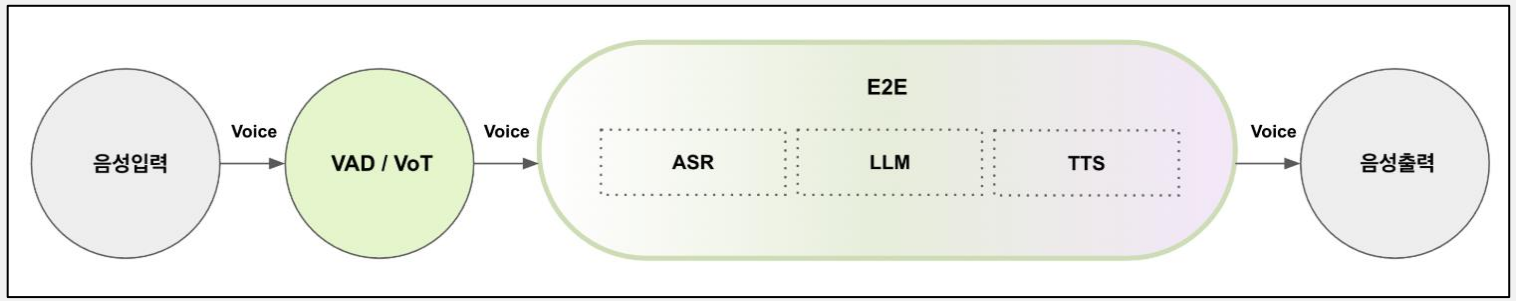
Case1: 에이닷 음성 모드

- 에이전트 음성모드로 QA 수행
- Pipeline 방식 기반으로 서비스 중
  - ASR → LLM → TTS 각 개별 모델의 input/output을 연계하여 동작
  - 개별 처리 단계(예: 음성 인식 → 텍스트 변환 → 자연어 처리 → 응답 생성)로 나누어 각 단계별로 별도의 모델이나 시스템이 처리하는 구조



Case2: 에이닷 전화 Agent

- 전화 대신걸기/대신받기 서비스
- GPT Realtime API기반 서비스 Prototype 중
  - 입력부터 출력까지 전체 과정을 하나의 Model로 처리



I. 에이닷 Voice Agent 현황

Usecase간 비교

	CASE1. 에이닷 음성모드	CASE2. 에이닷 전화 Agent (+상용 모델)
아키텍쳐	파이프라인 방식	E2E Model 방식
개발 방안	현재 Agent 에서 음성 분기를 통해, 음성 모드 전용으로 개발	전화망 연계한 실시간성 전화 대화 처리
정보의 입출력	각 단계별 입력 별도 처리	모든 입력을 통합적으로 한번에 처리
장점	• 다양한 액션을 수행 가능: 부가 서비스로의 연동 및 확장 • 구성 모델별 개별 개발	• 빠른 latency • 맥락, 억양, 감정 풍부하게 반영 가능
단점	• 복잡한 구조 • 지연 시간	• 비용 • 한국어 품질 부족 • LLM 직접 제어 불가로 대화 제어의 한계
업계 동향	Elevenlabs	GPT Realtime API, Gemini 2.5

자연스러운 실시간 대화를 위해서는 E2E Model 기반 Voice Agent 개발 필요
- Realtime API 대체할 수 있는 자체 솔루션 개발 필요
- 한국어 음성 기반 E2E 모델 파인 튜닝 필요

II. Voice Agent 기능별 요구사항

대화 Control 기능 요구사항

	기능	설명
Voice Agent 특화	VAD (Voice Activity Detection)	• **사용자 음성의 시작과 끝을 검출하는 기능** ◦ input audio의 시작과 끝 event message 확인 ◦ 실제 발화(Voice of Activity, VoT) 구간만 추출해 음성 신호의 전처리(불필요한 무음,노이즈 제거, 볼륨 정규화 등) ◦ 음성 입력 중 실제 발화 구간만을 감지해, 시스템 리소스 낭비를 줄이고 실시간성을 높임 ◦ VoT(Voice of Activity) 기술로 사용자가 말을 멈추는 순간을 감지, 자동으로 인식 구간을 종료
	Turn-taking	• **대화의 주고받음을 관리하는 기능** ◦ 사용자가 질문을 하면 Agent가 적절한 반응을 보이고, 대화가 원활하게 이어지는지 확인
	Barge-in	• **사용자가 대화 중에 중간에 끼어들 수 있는 기능** ◦ Barge-in 발생 시, 재생 중인 오디오 렌더링이 중단되고 새로운 대화가 진행되는지 확인
	Proactive Engagement	• **사용자의 상태나 대화 흐름을 고려하여 능동적으로 대화를 처리하는 기능** ◦ 사용자가 3초 이상 Silence 상태로 있으면 Agent가 능동적으로 발화를 유도하는지 확인 ◦ 침묵, 반복 질문, 무반응, 긴급상황, 대화 종료 등 대화 상태 판단 ◦ 상태에 따른 능동적 대화 처리, 발화 유도 멘트, 통화 종료 처리
	Security and Privacy	• **사용자의 개인정보 보호와 보안을 고려한 대화 처리 기능** ◦ 개인정보에 대해 적절하게 대응하는지 확인 ◦ 개인정보보호 정책(공개 범위 및 레벨)에 따른 개인정보 처리 ◦ 설정된 대화 범위 외 답변 제한 제어
	Conversation Context	• **대화의 전체 맥락을 제어하는 기능** ◦ Session의 Instruction Prompting 통해 대화 맥락이 제어되는지 확인 ◦ 대화 맥락 판단 및 이에 따른 Conversation Session 관리
Agent 공통	Memory	• **대화 중에 공유된 정보 및 상호작용을 기억하는 기능** ◦ 대화의 내용을 기억하고, 이를 바탕으로 자연스럽게 대화를 이어가는지 확인
	Personalized Conversation	• **사용자의 개인 정보를 기반으로 대화하는 기능** ◦ 사용자의 정보(개인화 메모리) 를 기반으로 개인화된 응답
	Function_Call Control	• **Function_Call 호출 기능** ◦ Function 정의, 파라미터 전달, 함수 호출 여부, 응답 값 확인 ◦ 개인화 메모리와 연계 Function_Call 개발 ◦ 대화 제어를 위한 Function_Call 개발 ◦ 필요한 요청에서만 Function_Call이 포함되도록 Latency 최적화 방안 필요
	Hallucination Control	• **Hallucination 발생 제어** ◦ Prompting을 통해 Hallucination 발생 제어를 시도하였으나 Hallucination 발생하기도 함 ◦ (예시) 스케줄 정보 혼동, 사용자 취향 가상 생성, 시제 혼동 등

II. Voice Agent 기능별 요구사항

STT(Speech-to-Text) 요구사항

기능	설명
VAD(Voice Activity Detection)	• 사용자 음성의 시작 및 끝 구간 검출 • VAD Default 설정에 대한 threshold 값 제공 • VAD threshold와 오디오 RMS(root mean square) 간의 관계 테이블 제공 • 배경 소음이 많은 환경에서도 음성을 정확히 인식할 수 있도록 Noise Suppression 기능 적용
한국어 특화 지원 (다국어 지원)	• 다국어 지원 • 한국어의 발음, 억양 및 문법적 특성을 반영하여, 한국어에 최적화된 음성 인식 기능 제공 • 한국어와 다국어가 섞인 발화에 대해서도 인식

오디오 코덱	• Codec, 샘플링 주파수, Bit Depth
	  ◦ 예시)
	  ▪ 지원 코덱: AMR, OPUS, PCM, G.711 지원 (선호 순서 AMR/OPUS, PCM, G.711)
	  ▪ 샘플링 주파수: 16 kHz/8kHz (16kHz 선호)
	  ▪ Bit Depth: 16bit/8bit (16bit 선호)
	• 코덱 특성을 고려하여 음성인식 최적화

정량적 평가 지표	• 음성인식의 정확도를 측정할 수 있는 정량적 평가 지표 제공 • 한국어 CER(character error rate) 7% 이하

II. Voice Agent 기능별 요구사항

TTS(Text-to-Speech) 요구사항

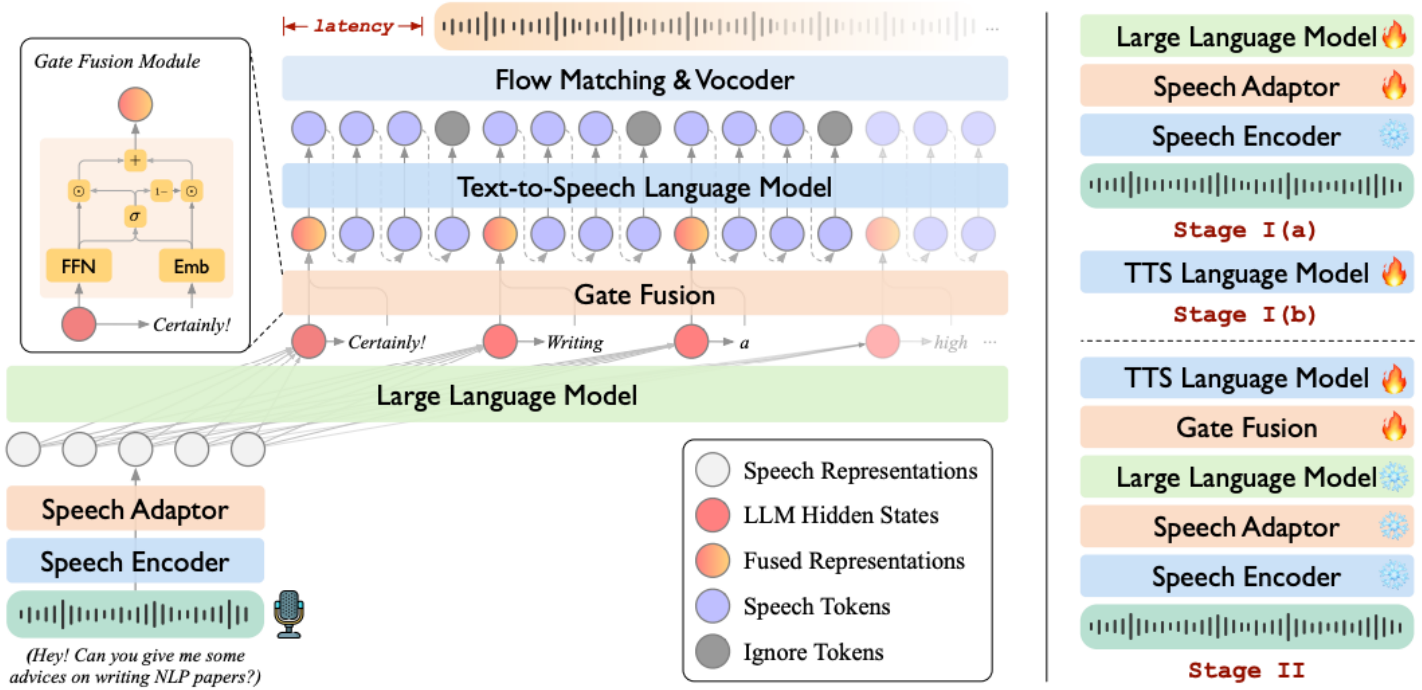
기능	설명
한국어 특화 지원 (다국어 지원)	• 한국어 발음, 억양, 문법적 특성을 고려하여 음성 합성 최적화 • 한국어 외에도 다양한 언어 지원
Voice Character Set 제공	• 최소 남여 각각 #명 이상 다양한 음성 제공 • 성별, 나이, 목소리 톤에 따른 Voice Character 특성 제공
자연스러운 음성 합성 기능 제공	• 자연스러운 발음과 억양 적용 • 대화의 맥락(Context) 에 맞는 감정을 표현할 수 있는 기능 제공 • TTS 속도 조절 기능 제공
Voice Cloning Tool 지원	• 사용자가 제공한 음성을 바탕으로 복제된 음성 생성 기능 제공 • Voice Cloning을 위한 최소 학습 오디오 길이 기준 제공 • Voice Cloning을 위한 최소 학습 오디오 품질 기준 제공 (통화 음성으로 Voice Cloning 지원)
대화 중 **Voice** 전환 기능 제공	• 에이전트 발화 Turn 단위로 Voice 전환 기능 지원
정성적/정량적 평가 지표 제공	• CMOS(Comparative Mean Opinion Score) • 동음이의어 발화, 동일단어 복수 발음 처리 정확도 지표

III. 개발 방안 논의

Voice Agent를 위한 E2E Model 공동 개발 제안

- Open source 모델 활용, 한국어 특화 모델로 개발
  - (예시) LLAMA-omni 모델 활용 등
- 전화서비스용 대화 파인튜닝 및 대화 control 기능 개발

(참고) LLAMA-omni2 모델 Architecture



감사합니다.